International Journal of Computer Networks and Distributed Computing
Vol. 3, Issue 2

Directional consistency and Relational consistency Concept Factorization for Text Clustering

V. Sunitha[1], Venkata Satya Ramesh Babu Batchu[2], U.Rakesh[3], K.Ramakrishna[4]

Assistant Professor, Department of CSE, Holy Mary Institute of Technology & Science, Hyderabad, India[1,2,3,4]

Abstract - This Paper is to extract the underlying concepts which are consistent with the low-dimensional manifold structure with the hope that this will facilitate further processing, such as clustering. Central to our approach is a graph model which captures the local geometry of the text submanifold. Thus, we call it Locally Consistent Concept Factoriaztion.The graph Laplacian, analogous to the Laplace-Beltrami operator on manifolds, can be used to smooth the text-to-concept mapping. Thus, the obtained concepts can well capture the intrinsic geometrical structure and the texts associated with similar concepts can be well clustered. The euclidean and manifold geometry is unified through a regularization framework where a regularization parameter controls their balance. Although the new approach is no longer optimal in the sense of reconstruction error in euclidean space, it has a better interpretation from manifold perspective. Moreover, like CF, our method also can be performed in RKHS which gives rise to nonlinear mappings.

Keywords:concept factorization, graphs Laplacian, manifold regularization, clustering, reproducing kernel Hilbert space

I.INTRODUCTION

In the last decade, matrix factorization-based approaches have attracted considerable attention in text clustering. When using matrix factorization-based methods, a text text is usually represented as a point in a high-dimensional linear space, each dimension corresponding to a term. Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered. Recent studies have shown that similarity can be measured more accurately in lower dimensional spaces, and thus the clustering performance can be enhanced. In particular, Nonnegative Matrix Factorization.and Concept Factorization have been applied to text clustering with impressive results. In general, the NMF problem is the following: given a nonnegative data matrix X, find reduced rank nonnegative matrices U and V so that UVT provides a good approximation to X. The column vectors of U can be thought of as basis vectors and V contains the coordinates. Previous studies have shown there is psychological and physiological evidence for parts-based representation in human brain. The nonnegative constraints in NMF lead to a partsbased representation because it allows only additive, not subtractive, combinations. The major limitation of NMF is that it is unclear how to effectively perform NMF in the transformed data space, e.g., reproducing kernel Hilbert space (RKHS).The euclidean and manifold geometry is unified through a regularization framework where a regularization parameter controls their balance. Although the new approach is no longer optimal in the sense of reconstruction error in euclidean space, it has a better interpretation from manifold perspective. Moreover, like CF, our method also can be performed in RKHS which gives rise to nonlinear mappings.

II.SYSTEM OVERVIEW

Electronic learning (e-Learning) refers to the application of information and communication technologies (e.g., Internet, multimedia, etc.) to enhance ordinary classroom teaching and learning. With the maturity of the technologies such as the Internet and the decreasing cost of the hardware platforms, more institutions are adopting e-Learning as a supplement to traditional instructional methods. In fact, one of the main advantages of e-Learning technology is that it can facilitate adaptive learning such that instructors can dynamically revise and deliver instructional materials in accordance with learners' current progress. In general, adaptive teaching and learning refers to the use of what is known about learners, a priori or through interactions, to alter how a learning experience unfolds, with the aim of improving learners' success and satisfaction. The current state-of the- art of e-Learning

technology supports automatic collection of learners' performance data (e.g., via online quiz).

However, few of the existing e-Learning technologies can support automatic analysis of learners' progress in terms of the knowledge structures they have acquired. In this paper, we illustrate a methodology of automatically constructing concept maps to characterize learners' understanding for a particular topic; thereby instructors can conduct adaptive teaching and learning based on the learners' knowledge structures as reflected on the concept maps. In particular, our concept map generation mechanism is underpinned by a context-sensitive text mining method and a fuzzy domain ontology extraction algorithm.

The notion of ontology is becoming very useful in various fields such as intelligent information extraction and retrieval, semantic Web, electronic commerce, and knowledge management. Although there is not a universal consensus on the precise definition of ontology, it is generally accepted that ontology is a formal specification of conceptualization.

Ontology can take the simple form of a taxonomy of concepts (i.e., light weight ontology), or the more comprehensive representation of comprising a taxonomy, as well as the axioms and constraints which characterize some prominent features of the real-world (i.e., heavy weight ontology). Domain ontology is one kind of ontology which is used to represent the knowledge for a particular type of application domain. On the other hand, concept maps are used to elicit and represent the knowledge structure such as concepts and propositions as perceived by individuals. Concept maps are similar to ontology in the sense that both of these tools are used to represent concepts and the semantic relationships among concepts.

However, ontology is a formal knowledge representation method to facilitate human and computer interactions and it can be expressed by using formal semantic markup languages such as RDF and OWL, whereas concept map is an informal tool for humans to specify semantic knowledge structure. Figure shows an example of the owl statements describing one of the fuzzy domain ontologies automatically generated from our system. It should be noted that we use the (rel) attribute of the <rdfs:comment> tag to describe the membership of a fuzzy relation (e.g., the super-class/sub-class relationship). We only focus on the automatic extraction of lightweight domain ontology in this paper. More specificially, the lightweight fuzzy domain ontology is used to generate concept maps to represent learners' knowledge structures.

With the rapid growth of the applications of e-Learning to enhance traditional instructional methods, it is not surprising to find that there are new issues or challenges arising when educational practitioners try to bring information technologies down to their classrooms. The situation is similar to the phenomenon of the rapid growth of the Internet and the World Wide Web (Web). The explosive growth of the Web makes information seekers become increasingly more difficult to find relevant information they really need.

membership degrees (classification) for each event with respect to the fuzzy concepts defined in the fuzzy ontology. The standard triangular membership function was used for the classification purpose.

The method discussed in this paper is a fully automatic fuzzy domain ontology discovery approach. There is no predefined fuzzy concepts and taxonomy of concepts, instead our fuzzy domain ontology extraction method will automatically discover the concepts and generate the taxonomy relations. In addition, there is no need to set the artificial threshold values for the triangular membership function, instead our membership function can automatically derive the membership values based on the lexico-syntactic and statistical features of the terms observed in a textual database. An ontology mining technique was proposed to extract patterns representing users' information needs. The ontology mining method consists of two parts: the top backbone and the base backbone. The former represents the relations between compound classes of the ontology. The latter indicates the linkage between primitive classes and compound classes.

It has been pointed out that the main challenge of automatic ontology extraction from textual databases is the removal of noisy concepts and relations. Based on this premise, our domain ontology extraction methodology in general and the concept map generation process in particular are designed to effectively filter the non-relevant concepts and concept relations from the concept space. Figure 2 depicts the proposed methodology of automatically generating concept maps from a collection of online messages posted to blogs, emails, chat rooms, Web pages, etc. The collection of messages is treated as a textual corpus. At the text parsing stage, our text parser will scan each message to analyze the lexico-syntactic elements embedded in the message. For instance, stop words such as "a, an, the" are removed from the message since these words appear in any contexts and they cannot provide useful information stop word file used in the SMART retrieval system. Different customizations is required for processing different kinds of texts. For example, we need to extend the SMART stop word file by including stop words such as "home", "contact", "web", "site", etc. for parsing Web pages.

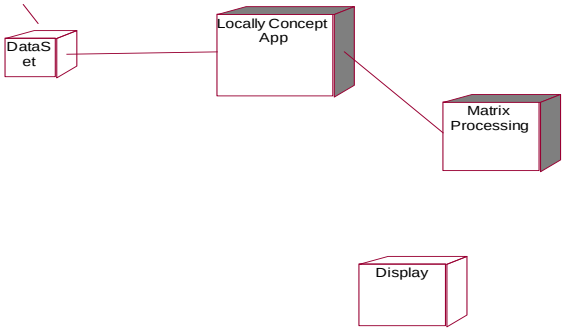


Fig 1: system design diagram for all modules

Lexical patterns are identified by applying Part-of-Speech (POS) tagging to the source texts. We develop our POS tagger based on the WordNet lexicon and the publicly available API (http://wordnet.princeton.edu/). For namedentity detection (e.g., people's names, organizations' names, etc.), we employ BBN's IdentiFinder. However, for the e- Learning application reported in this paper, we do not make use of the entity tags for concept extraction. We simply treat each named-entity as a noun for subsequent linguistic pattern mining.

A text windowing process will be conducted by scanning adjacent tokens within a pre-defined window size of 5 to 10 words from left to right over all the texts. At the end of the windowing process, an information theoretic measure is applied to compute the co-occurrence statistics between the targeting linguistic patterns and other tokens appearing in the same text window across the corpus. Thereby, context vectors can be created to describe the semantic of the extracted concepts.

Part of the Semantic Web vision is to provide web-scale access to semantically described content. In particular, this implies understanding users' information needs accurately enough to allow for retrieving a precise answer using semantic technologies.

After the tagging process, each token is stemmed according to the Porter stemming algorithm. During the concept extraction stage (Section V-A), certain linguistic patterns are ignored to reduce the generation of noisy concepts. For example, ontology engineers or instructors in the case of e- Learning application, will specify the mining focus on certain linguistic patterns such as "Noun Noun", "Adjective Noun", "Verb Noun", etc. The text mining program will then focusing on finding the term association information and collecting the statistical data for those patterns only. Not only does it reduce the generation of noisy concepts but also improve the computational efficiency of our ontology extraction process.

Currently, most web search engines are however based on purely statistical techniques. While they are not able to figure out the meaning of a query, they can provide answers by returning the statistically most appropriate answer to a user's query—based on some measures for computing similarity in vector space. Information Retrieval (IR) techniques applied to the Web have gained a reasonable degree of maturity which is clearly corroborated by the success of search engines such as Google, Yahoo and the like. These search engines are in fact providing a baseline quite difficult to outperform.

Search queries are the articulation of a person's information goals. People employ a mixture of search and navigation strategies to satisfy these goals. In laboratory studies, participants can be given known search goals or probed about their own information goals. In large-scale log studies, information goals must be inferred from patterns of user interactions. One approach to inferring searchers' information goals is to consider the patterns of pages viewed in sessions as well as the dwell times on pages as implicit indicators of interest.

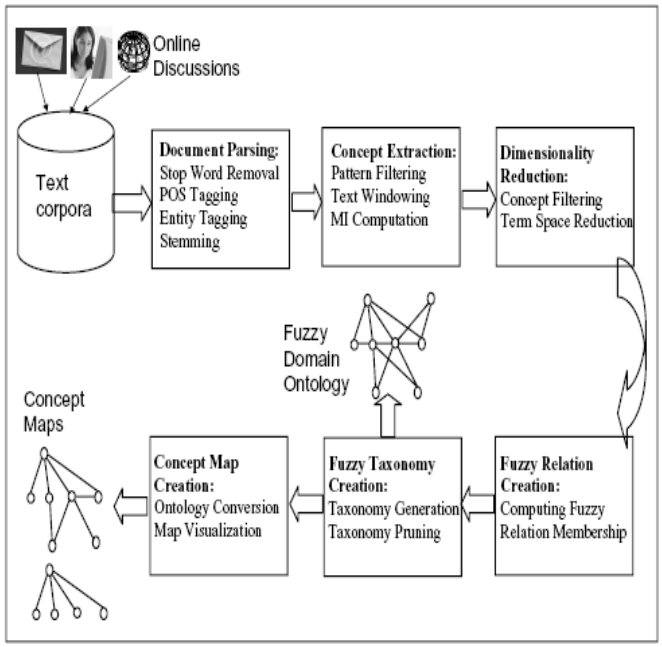


Fig. 2. A Framework for Automatic Concept Map Generation

The Semantic Web aims to use semantics in the retrieval process, where the semantics is captured in ontologies or at the very least in concept hierarchies. The task then is to find pairs of concepts from different meta-data schemas that have an equivalent meaning, a problem known as ontology matching. This problem has been extensively studied in the Semantic Web and elsewhere, for recent survey papers. However, in many realistic domains, it is impossible to give precise concept definitions, and consequently no crisp notion of concept equivalence exists.

Below we will illustrate this in the music domain (an important commercial domain on the Web), where musical genres are inherently imprecise. Such imprecision is a fundamental aspect of many other domains as well. Ontology matching must then be redefined to finding a concept with the closest meaning in the other schema when an equivalent one does not exist. We then require mechanisms that are able to find approximate correspondences rather than exact ones.

Before moving to the technical part of the paper, we first briefly introduce the domain of musical genres, and will argue why this is an appropriate domain for investigating techniques for approximate ontology mapping.

III.THE WORKING PRINCIPLE

Concept Factorization:
In this module Nonnegative Matrix Factorization is used as a matrix factorization algorithm that focuses on the analysis of data matrices whose elements are nonnegative. Given a nonnegative data matrix, each column is a sample vector. NMF aims to find two nonnegative matrices which minimize the objective function. Although the objective function O is convex, it is not convex in both variables together. Therefore, it is unrealistic to expect an algorithm to find the global minimum we use iterative algorithm where uk is the kth column vector of U. Thus, each data vector xj is approximated by a linear combination of the columns of U, weighted by the components of V. Therefore, U can be regarded as containing a basis that is optimized for the linear approximation of the data in X. This can be regarded as the new representation of each data point in the new basis U. Since relatively fewer basis vectors are used to represent many data vectors, good approximation can only be achieved if the basis vectors discover structure that is latent in the data..

Objective Function: This module uses the CF and tries to find a basis that is optimized for the linear approximation of the data. The jth row of matrix V, can be regarded as the new representation of each data point in the new basis. One might hope that knowledge of the geometric structure of the data can be exploited for better discovery of this basis. A natural assumption here could be that if two data points xj; xs are close in the intrinsic geometry of the data distribution, then zj and zs, the representations of this two points in the new basis, are also close to each other.

Multiplicative Algorithm
In this module we use the objective function O which is not convex in both W and V together. Therefore, it is unrealistic to expect an algorithm to find the global minimum of O. In the following, we introduce an iterative algorithm which can achieve a local minimum. Regarding these two updating rules, by computing the non-increasing under the updating rules. The objective function is invariant under these updates if and only if W and V are at a stationary point. The updating rules of W and V converge and the final solution will be a local optimum. For the objective function of CF, it is easy to check that if Wand V are the solution then will also form a solution for any positive diagonal matrix D. here w is the column vector of W. The matrix V will be adjusted accordingly.

Gradient Descent
In this module we use algorithm for minimizing the objective function which is gradient descent. The step size parameters are used which are sufficiently small, and the updates will reduce the complexity. Generally speaking, it is relatively hard to set these step size parameters while still maintaining the non negativity of Matrix. However, with the special form of the partial derivatives, we have the multiplicative updating rules which are special cases of gradient descent with automatically step size parameter selection. The advantage of multiplicative updating rules is the guarantee the non-negativity of W and V vectors.

Negative Data Matrices
This module creates and woks on K which is nonnegative. In the case that the data matrix has negative values, it is possible that the K has negative entries. In this module, we will introduce a general algorithm which can be applied for any case. A is a symmetric positive definite matrix and b is an arbitrary m-dimensional vector. we can easily see that the objective function is a quadratic form and we only need to identify the corresponding A and b in the objective function.

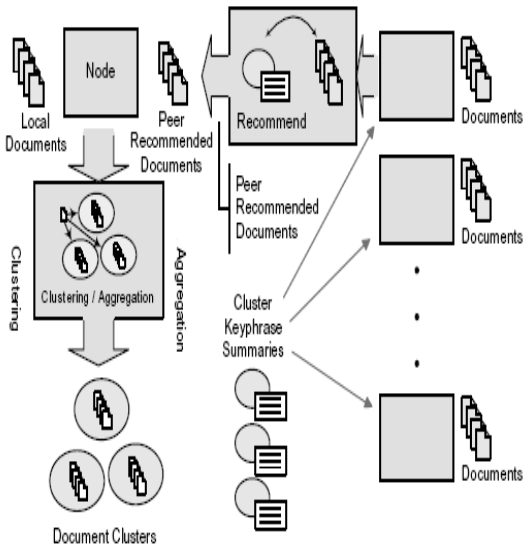


Fig3: System Architecture

IV.IMPLEMENTATION OF SYSTEM

A.Pseudo for Concept Factorization:

```
public InputStream is;
public OutputStream os;
   public SocketConnection sc;
   public ServerSocketConnection scn;
    int isel = choiceGroup1.getSelectedIndex();
                if(isel==0)
              {
                 Thread t = new Thread(this,"T1");
                 flag =1;
                 t.start();
              }
                  if(isel==1)
              {
                 getDisplay().setCurrent(get_SendForm());
              }
              if(isel==2)
              {
              getDisplay().setCurrent(get_ReceivePortForm());
              }
```

B.Pseudo for Objective Function:

```
if(flag==1)
      {
         try
         {
            ServerCall sc = new ServerCall();
         String key =
sc.sendRequest("http://localhost:9090/CentralAuthority/Ce
ntralAuth?message="+textField1.getString());
         System.out.println(key);
         }
         catch (Exception ex)
         {
            ex.printStackTrace();
         }
      }
```

C.Pseudo for Multiplicative Algorithm:

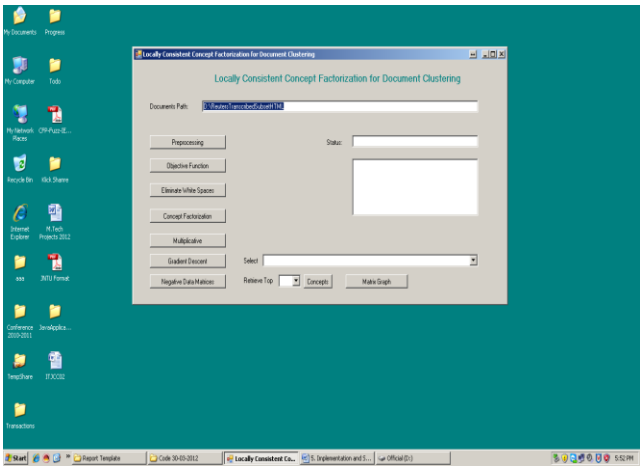
```
    scn =       (ServerSocketConnection)Connector.open
("socket://:"+t    extField7.getString());
         System.out.println(scn.toString());
         System.out.println("Waiting for Connection ");
         sc = (SocketConnection)scn.acceptAndOpen();
         System.out.println("Connection Accepted ");
         is = sc.openInputStream();
         os = sc.openOutputStream();
         StringBuffer sb;
         String str = "";
         int c=0;
         while (((c = is.read()) != -1))
               str += (char)c;
         System.out.println("Hello");
         textField5.setString(str);
         DES d2 = new DES();
                  d2.initialize();
                  String plain = d2.decrypt(cipher);
String plain = d2.decrypt(str);
System.out.println(plain);
 textField6.setString(plain);
 System.out.println(str);
```

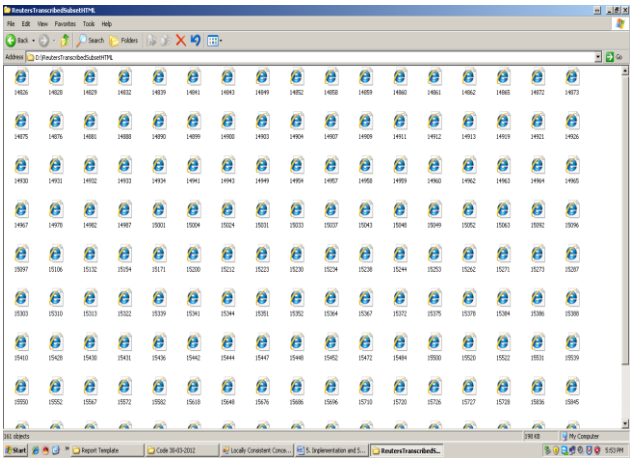
D.Pseudo for Gradient Descent:

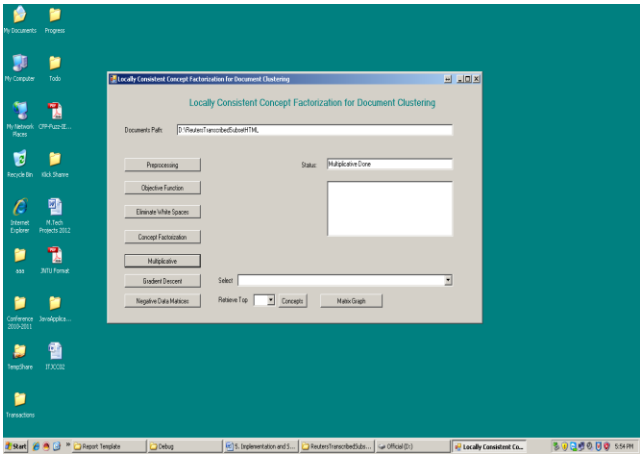
```
Socket Connection sc1 =
(SocketConnection)Connector.open("socket://localhost:"+t
extField3.getString());
         System.out.println("Connected ");
         InputStream is1 = sc1.openInputStream();
         Output Stream os1  = sc1.openOutputStream();
         DES d1 = new DES();
 d1.initialize();
String cipher = d1.encrypt(textField4.getString());
System.out.println(cipher);
       os1.write(d1.hexcipher.getBytes());
         os1.flush();
         os1.close();
         sc1.close();
```

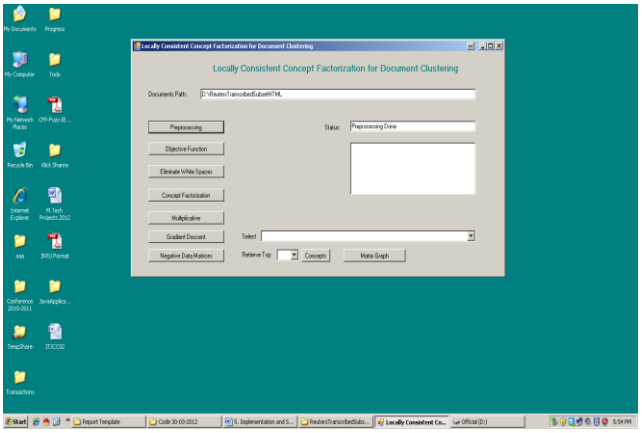
V.EXPERIMENTAL RESULTS

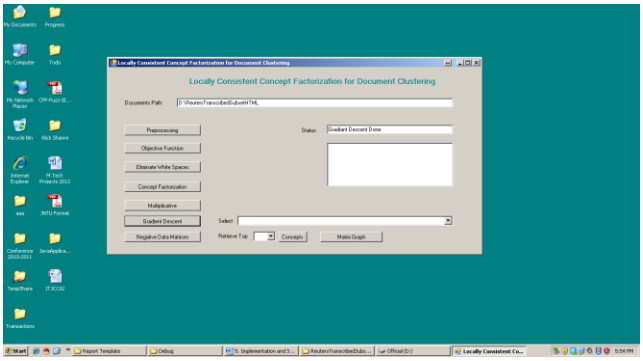
The concept of this paper is implemented and different results are shown below.

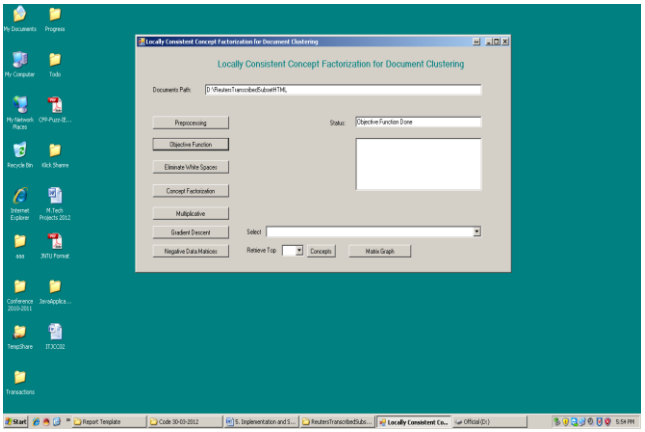


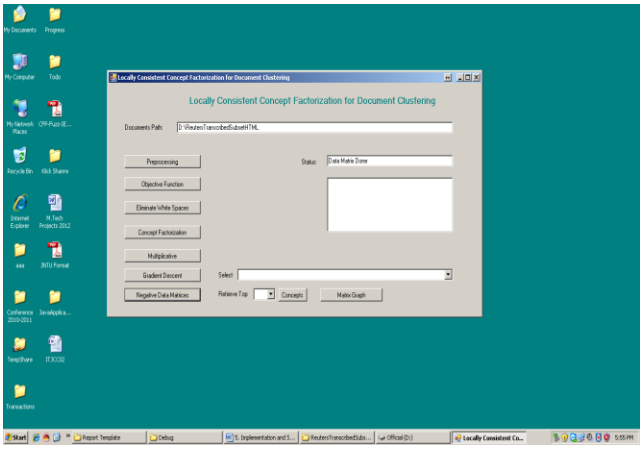












VI.CONCLUSION

In this paper propose a new approach to extract the text concepts which are consistent with the manifold geometry such that each concept corresponds to a connected component. Central to our approach is a graph model which captures the local geometry of the text submanifold.presented to extract the underlying concepts which are consistent with the low-dimensional manifold structure with the hope that this will facilitate further processing, such as clustering. Central to our approach is a graph model which captures the local geometry of the text

sub manifold. Thus, we call it Locally Consistent Concept Factoriaztion.The graph Laplacian, analogous to the Laplace-Beltrami operator on manifolds, can be used to smooth the text-to-concept mapping. Thus, the obtained concepts can well capture the intrinsic geometrical structure and the texts associated with similar concepts can be well clustered.

.

REFERENCES

[1] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," Advances in Neural Information Processing Systems 14, pp. 585-591, MIT Press, 2001.
[2] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization:.
[3] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N.L. Roux, and M. Ouimet, "Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering,.
[4] J.-P. Brunet, P. Tamayo, T.R. Golub, and J.P. Mesirov, "Metagenes and Molecular Pattern Discovery Using Matrix Factorization," Proc. Nat'l Academy of Sciences USA vol. 101, no. 12, pp. 4164-4169, 2004.
[5] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. harshman, "Indexing by Latent Semantic Analysis," J. Am. Soc. of Information Science, vol. 416, pp. 391-407, 1990.
[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Royal Statistical Soc., Series B (Methodological), vol. 391, pp. 1-38, 1977.
[7] I.S. Dhillon, Y. Guan, and B. Kulis, "Kernel K-Means: Spectral Clustering and Normalized Cuts," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '04), pp. 551-556, 2004.
[8] J. Kivinen and M.K. Warmuth, "Additive versus Exponentiated Gradient Updates for Linear Prediction," Proc. 27th Ann. ACM Symp. Theory of Computing (STOC '95), pp. 209-218, 1995.
[9] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization," Nature, vol. 401, pp. 788-791, 1999.
[10] Deng Cai, Member, IEEE, Xiaofei He, Senior Member, IEEE, and Jiawei Han, Fellow, IEEE Locally Consistent Concept Factorization for Document Clustering.